

Weighted-indexed semi-Markov models for modeling financial returns

Guglielmo D'Amico

Dipartimento di Scienze del Farmaco, Facoltà di Farmacia, Università "G. D'Annunzio" di Chieti-Pescara, 66013 Chieti, Italy

Filippo Petroni

Dipartimento di Scienze Economiche e Aziendali, Facoltà di Economia, Università degli studi di Cagliari, 09123 Cagliari, Italy

Abstract. In this paper we propose a new stochastic model based on a generalization of semi-Markov chains to study the high frequency price dynamics of traded stocks. We assume that the financial returns are described by a weighted indexed semi-Markov chain model. We show, through Monte Carlo simulations, that the model is able to reproduce important stylized facts of financial time series as the first passage time distributions and the persistence of volatility. The model is applied to data from Italian and German stock market from 1 January 2007 until the end of December 2010.

1. Introduction

Semi-Markov processes (SMP) are a wide class of stochastic processes which generalize at the same time both Markov chains and renewal processes. The main advantage of SMP is that they allow the use of whatever type of waiting time distribution for modeling the time to have a transition from one state to another one. On the contrary, Markovian models have constraints on the distribution of the waiting times in the states which should be necessarily represented by memoryless distributions (exponential or geometric for continuous and discrete time cases respectively). This major flexibility has a price to pay: the parameters to be estimated are more numerous.

Semi-Markov processes (SMP) generalizes also non-Markovian models based on continuous time random walks extensively used in the econophysics community, see for example [1, 2]. SMP have been used to analyze financial data and to describe different problems ranging from credit rating data modeling [3] to the pricing of options [4, 5].

With the financial industry becoming fully computerized, the amount of recorded data, from daily close all the way down to tick-by-tick level, has exploded. Nowadays, such tick-by-tick high-frequency data are readily available for practitioners and researchers alike [6, 7]. It seemed then natural to us trying to verify the semi-Markov hypothesis of returns on high-frequency data, see [8]. In [8] we proposed a semi-Markov model showing its ability to reproduce some stylized empirical facts such for example the absence of autocorrelations in returns and the gain/loss asymmetry. In that paper we showed also that the autocorrelation in the square of returns is higher with respect to the Markov model. Unfortunately this autocorrelation was still too small compared to the empirical one. In order to overcome the problem of low autocorrelation, in another paper [9] we proposed an indexed semi-Markov model for price return. More precisely we assumed that the intraday returns (up to one minute frequency) are described by a discrete time homogeneous semi-Markov process where we introduced a memory index which takes into account the periods of different volatility in the market. It is well known that the market volatility is autocorrelated, then periods of high (low) volatility may persist for long time. We made the hypothesis that the kernel of the semi-Markov process depend on which level of volatility the market is at that time. It is to be remarked that the weighted memory index is a stochastic process which do depend on the same Markov Renewal Chain (J_n, T_n) to which the semi-Markov chain is associated. Then, in our model, the high autocorrelation is obtained endogenously without introducing external or latent auxiliary stochastic processes. To improve further our previous results, in this work, we propose an exponentially weighted index which will be described in the following.

The database used for the analysis is made of high frequency tick-by-tick price data from all the stock in Italian and German stock market from first of January 2007 until end of December 2010. From prices we then define returns at one minute frequency.

The plan of the paper is as follows. In Section 2 we define the weighted indexed semi-Markov chain model with memory and we explain how to perform a Monte Carlo

simulation of its trajectory. In Section 3, we present the empirical results deriving from the application of our model to real stock market data. Finally, in Section 4 we present our conclusion.

2. The Weighted-Indexed Semi-Markov Model

In this section we propose a generalization of the semi-Markov process that is able to represent higher-order dependencies between successive observations of a state variable. One way to increase the memory of the process is by using high-order semi-Markov processes as defined in [10] and more recently revisited and extended in a discrete time framework in [11]. A more parsimonious model has been defined by [12] and it is showed that it describes appropriately important empirical regularities of financial time series. In this paper we propose a further improvement of the indexed semi-Markov chain model proposed in reference [12] named Weighted-Indexed Semi-Markov Chain (WISMC) model which allows the possibility of reproducing long-term dependence in the square of stock returns in a very efficient way.

Let us assume that the value of the financial asset under study is described by the time varying asset price $S(t)$. The return at time t calculated over a time interval of length 1 is defined as $\frac{S(t+1)-S(t)}{S(t)}$. The return process changes value in time, then we denote by $\{J_n\}_{n \in \mathbb{N}}$ the stochastic process with finite state space $E = \{1, 2, \dots, s\}$ and describing the value of the return process at the n -th transition.

Let us consider the stochastic process $\{T_n\}_{n \in \mathbb{N}}$ with values in \mathbb{N} . The random variable T_n describes the time in which the n -th transition of the price return process occurs.

Let us consider also the stochastic process $\{U_n^\lambda\}_{n \in \mathbb{N}}$ with values in \mathbb{R} . The random variable U_n^λ describes the value of the index process at the n -th transition.

In reference [9] the process $\{U_n\}$ was defined as a reward accumulation process linked to the Markov Renewal Process $\{J_n, T_n\}$; in [9] the process $\{U_n\}$ was defined as a moving average of the reward process. Here, motivated by the application to financial returns, we consider a more flexible index process defined as follows:

$$U_n^\lambda = \sum_{k=0}^{n-1} \sum_{a=T_{n-1-k}}^{T_{n-k}-1} f(J_{n-1-k}, a, \lambda), \quad (1)$$

where $f : E \times \mathbb{N} \times \mathbb{R} \rightarrow \mathbb{R}$ is a Borel measurable bounded function and U_0^λ is known and non-random.

The process U_n^λ can be interpreted as an accumulated reward process with the function f as a measure of the weighted rate of reward per unit time. The function f depends on the current time a , on the state J_{n-1-k} visited at current time and on the parameter λ that represents the weight.

In next section a specific functional form of f will be selected in order to produce a real data application.

To construct the WISMC model we have to specify a dependence structure between the variables. Toward this end we adopt the following assumption:

$$\begin{aligned} \mathbb{P}[J_{n+1} = j, T_{n+1} - T_n \leq t | \sigma(J_h, T_h, U_h^\lambda), h = 0, \dots, n, J_n = i, U_n^\lambda = v] \\ = \mathbb{P}[J_{n+1} = j, T_{n+1} - T_n \leq t | J_n = i, U_n^\lambda = v] := Q_{ij}^\lambda(v; t), \end{aligned} \quad (2)$$

where $\sigma(J_h, T_h, U_h^\lambda)$, $h \leq n$ is the natural filtration of the three-variate process.

The matrix of functions $\mathbf{Q}^\lambda(v; t) = (Q_{ij}^\lambda(v; t))_{i,j \in E}$ has a fundamental role in the theory we are going to expose, in recognition of its importance, we call it *weighted-indexed semi-Markov kernel*.

The joint process (J_n, T_n) depends on the process U_n^λ , the latter acts as a stochastic index. Moreover, the index process U_n^λ depends on (J_n, T_n) through the functional relationship (1).

Observe that if

$$\mathbb{P}[J_{n+1} = j, T_{n+1} - T_n \leq t | J_n = i, U_n^\lambda = v] = \mathbb{P}[J_{n+1} = j, T_{n+1} - T_n \leq t | J_n = i]$$

for all values $v \in \mathbb{R}$ of the index process, then the weighed indexed semi-Markov kernel degenerates in an ordinary semi-Markov kernel and the WISMC model becomes equivalent to classical semi-Markov chain model as presented for example in [13] and [14].

The triple of processes $\{J_n, T_n, U_n^\lambda\}$ describes the behaviour of the system only in correspondence of the transition times T_n . To describe the behavior of our model at whatever time t which can be a transition time or a waiting time, we need to define additional stochastic processes.

Given the three-dimensional process $\{J_n, T_n, U_n^\lambda\}$ and the weighted indexed semi-Markov kernel $\mathbf{Q}^\lambda(v; t)$, we define by

$$\begin{aligned} N(t) &= \sup\{n \in \mathbb{N} : T_n \leq t\}; \\ Z(t) &= J_{N(t)}; \\ U^\lambda(t) &= \sum_{k=0}^{N(t)-1+\theta} \sum_{a=T_{N(t)+\theta-1-k}}^{(t \wedge T_{N(t)+\theta-k})-1} f(J_{N(t)+\theta-1-k}, a, \lambda), \end{aligned} \quad (3)$$

where $\theta = 1_{\{t > T_{N(t)}\}}$.

The stochastic processes defined in (3) represent the number of transitions up to time t , the state of the system (price return) at time t and the value of the index process (weighted moving average of function of price return) up to t , respectively. We refer to $Z(t)$ as a weighted indexed semi-Markov process.

The process $U^\lambda(t)$ is a generalization of the process U_n^λ where time t can be a transition or a waiting time. It is simple to realize that if $t = T_n$ we have that $U^\lambda(t) = U_n^\lambda$.

Let

$$p_{ij}^\lambda(v) := \mathbb{P}[J_{n+1} = j | J_n = i, U_n^\lambda = v],$$

be the transition probabilities of the embedded indexed Markov chain. It denotes the probability that the next transition is in state j given that at current time the process entered in state i and the index process is equal to v . It is simple to realize that

$$p_{ij}^\lambda(v) = \lim_{t \rightarrow \infty} Q_{ij}^\lambda(v; t). \quad (4)$$

Let $H_i^\lambda(v; \cdot)$ be the sojourn time cumulative distribution in state $i \in E$:

$$H_i^\lambda(v; t) := \mathbb{P}[T_{n+1} - T_n \leq t | J_n = i, U_n^\lambda = v] = \sum_{j \in E} Q_{ij}^\lambda(v; t). \quad (5)$$

It expresses the probability to make a transition from state i with sojourn time less or equal to t given the indexed process is v .

The conditional waiting time distribution function G expresses the following probability:

$$G_{ij}^\lambda(v; t) := \mathbb{P}[T_{n+1} - T_n \leq t | J_n = i, J_{n+1} = j, U_n^\lambda = v]. \quad (6)$$

It is simple to establish that

$$G_{ij}^\lambda(v; t) = \begin{cases} \frac{Q_{ij}^\lambda(v; t)}{p_{ij}^\lambda(v)} & \text{if } p_{ij}^\lambda(v) \neq 0 \\ 1 & \text{if } p_{ij}^\lambda(v) = 0. \end{cases} \quad (7)$$

In the papers [9] and [8] explicit renewal-type equations were given to describe the probabilistic behavior of the indexed semi-Markov chain. It is possible to derive similar results for the WISMC model but here we prefer to not report these results applied to our model because, in the implementation of the model given in the next section, we follow a Monte Carlo simulation based approach. Monte Carlo methods are very useful for simulating the system behavior and represent a way of generate replicable results. In the following we give a Monte Carlo algorithm in order to simulate a trajectory of a given WISMC in the time interval $[0, T]$. The algorithm consists in repeated random sampling to compute successive visited states of the random variables $\{J_0, J_1, \dots\}$, the jump times $\{T_0, T_1, \dots\}$ and the index values $\{U_0^\lambda, U_1^\lambda, \dots\}$ up to the time T .

The algorithm consists of 5 steps:

- 1) Set $n = 0$, $J_0 = i$, $T_0 = 0$, $U_0^\lambda = v$, horizon time = T ;
- 2) Sample J from $p_{J_n, \cdot}^\lambda(U_n^\lambda)$ and set $J_{n+1} = J(\omega)$;
- 3) Sample W from $G_{J_n, J_{n+1}}^\lambda(U_n^\lambda, \cdot)$ and set $T_{n+1} = T_n + W(\omega)$;
- 4) Set $U_{n+1}^\lambda = \sum_{k=0}^n \sum_{a=T_{n-k}}^{T_{n+1-k}-1} f(J_{n-k}, a, \lambda)$;
- 5) If $T_{n+1} \geq T$ stop
else set $n = n + 1$ and go to 2).

3. Empirical results

In the following we show as our model performs comparing its statistical features and those of real data returns. The comparison is done by means of Monte Carlo simulations according to the algorithms described in the previous section.

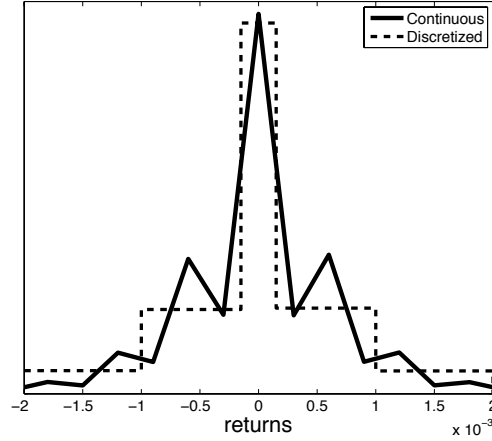


Figure 1. Discretization of returns

For our analysis we choose 4 stocks from two databases of tick-by-tick quotes of real stocks from the Italian Stock Exchange (“Borsa Italiana”) and the German Stock Exchange (“Deutsche Börse”). The chosen stocks are ENI and FIAT from the Italian database and Allianz and VolksWagen from the German database. The period used goes from January 2007 to December 2010 (4 full years). The data have been re-sampled to have 1 minute frequency. The number of returns analyzed is then roughly 500×10^3 for each stock.

To be able to model returns has a semi-Markov process the state space has been discretized. In the 4 examples shown in this work, we discretized returns into 5 states chosen to be symmetrical with respect to returns equal zero and to keep the shape of the distribution unchanged. Returns are in fact already discretized in real data due to the discretization of stock prices which is fixed by each stock exchange and depends on the value of the stock. Just to make an example, in the Italian stock market for stocks with value between 5.0001 and 10 euros the minimum variation is fixed to 0.005 euros (usually called tick). We then tried to remain as much as possible close to this discretization. In Figure 1 we show an example of the discretization of the returns of one of the analyzed stocks. The model described in the previous section requires the specification of a function f in the definition of the weighted index U_n^λ in (1). Let us briefly remind that volatility of real market is long range positively autocorrelated and then clustered in time. This implies that, in the stock market, there are periods of high and low volatility. Motivated by this empirical facts we suppose that also the transition probabilities depends on whether the market is in a high volatility period or in a low one. In a previous work [9], for simplicity reason, we used a moving average of the squares of returns as the index variable U . In that case we imposed that the index depended only on a memory m which was the number of transitions in the past used for the moving average. In this work we decided to use a more appropriate expression for f . We use an exponentially weighted moving average (EWMA) of the squares of returns which as

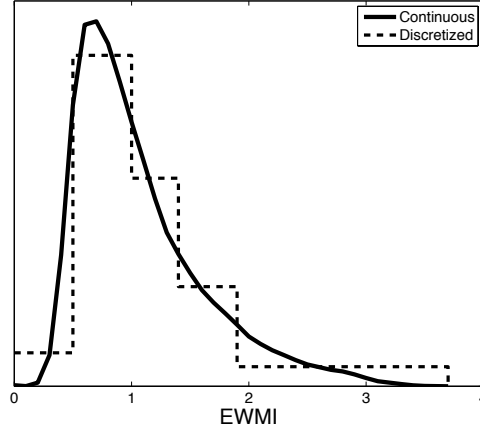


Figure 2. Discretization of index values

the following expression:

$$f(J_{n-1-k}, a, \lambda) = \frac{\lambda^{T_n-a} J_{n-1-k}^2}{\sum_{k=0}^{n-1} \sum_{a=T_{n-1-k}}^{T_{n-k}-1} \lambda^{T_n-a}} \quad (8)$$

and consequently the index process becomes

$$U_n^\lambda = \sum_{k=0}^{n-1} \sum_{a=T_{n-1-k}}^{T_{n-k}-1} \left(\frac{\lambda^{T_n-a} J_{n-1-k}^2}{\sum_{k=0}^{n-1} \sum_{a=T_{n-1-k}}^{T_{n-k}-1} \lambda^{T_n-a}} \right). \quad (9)$$

The index U^λ was also discretized into 5 states of low, medium low, medium, medium high and high volatility. An example of the discretization used in the analysis is shown in Figure 2.

A very important feature of stock market data is that, while returns are uncorrelated and show an i.i.d. like behavior, their square or absolute values are long range correlated. It is very important that theoretical models of returns reproduce this features. We then tested our model to check whether it is able to reproduce such behavior. Given the presence of the parameter λ in the index function, we tested the autocorrelation behavior as a function of λ . Note that in the definition of the index variable the EWMA is performed over all the previous square of returns each with its weight. Before summing over all past returns we decided to check whether a better memory time m exists. For this reason we checked our model also against this other parameter. With this choice formula (9) takes the form:

$$U_n^\lambda(m) = \sum_{k=n-m}^{n-1} \sum_{a=T_{n-1-k}}^{T_{n-k}-1} \left(\frac{\lambda^{T_n-a} J_{n-1-k}^2}{\sum_{k=n-m}^{n-1} \sum_{a=T_{n-1-k}}^{T_{n-k}-1} \lambda^{T_n-a}} \right). \quad (10)$$

We remind the definition of the autocorrelation function: if Z indicates returns, the time lagged (τ) autocorrelation of the square of returns is defined as

$$\Sigma(\tau) = \frac{Cov(Z^2(t+\tau), Z^2(t))}{Var(Z^2(t))} \quad (11)$$

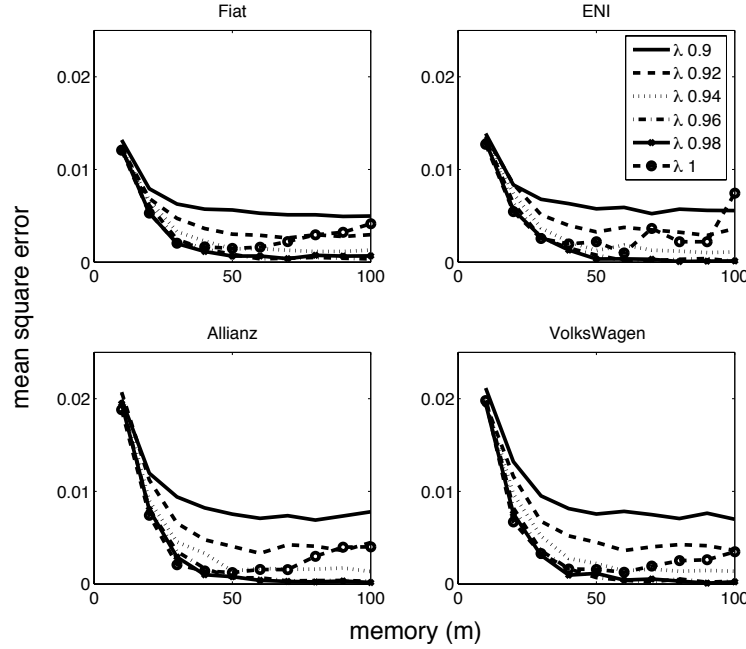


Figure 3. Mean square error between autocorrelation functions from real and simulated data as functions of m and for different values of λ .

We estimated $\Sigma(\tau)$ for real data and for returns time series simulated with different values of the memory time m and the weights λ . The time lag τ was made to run from 1 minute up to 100 minutes. Note that to be able to compare results for $\Sigma(\tau)$ each simulated time series was generated with the same length as real data. In Figure 3 we show the mean square error between $\Sigma(\tau)$ obtained from real and simulated returns (using definition (10) for the index process) for the four stocks analyzed and for different m and λ . Let us make some considerations on the results shown in Figure 3: m should be chosen as big as possible and then definition (9) is appropriate as far as λ is chosen less than 1, in fact, in this last case definition (9) becomes equivalent to a moving average without weights and results presented in [9] hold for m . In Figure 4 we show again the mean square error but only as a function of the weights λ then using definition (9) for the index process. We can notice that the behavior is very similar for the different analyzed stocks even if the best value for λ is not the same for all of them. As it is possible to see the best values of λ for the stocks Fiat, Eni, Allianz and VolksWagen are 0.96, 0.97, 0.97 and 0.98, respectively.

The comparison between the autocorrelations for the best values of λ for each stock and real data is shown in Figure 5. This figure shows that real and synthetic data have almost the same autocorrelation function for the square of returns.

We tested our model also to verify if it is able to reproduce the feature shown by real data regarding the first passage time (fpt) distribution [8, 15, 16]. Let us remind here the definition of fpt: the fpt for an investment made at time t at price $S(t)$ is defined as the time interval $\tau = t' - t$, $t' > t$ where the relation $S(t + \tau)/S(t) \geq \rho$ is fulfilled for

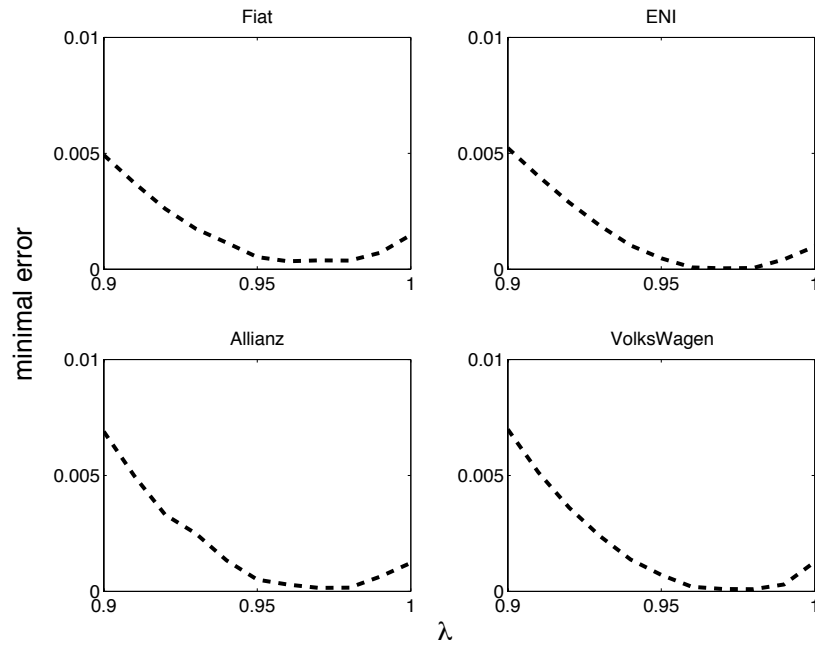


Figure 4. Mean square error between autocorrelation functions from real and simulated data as functions of λ .

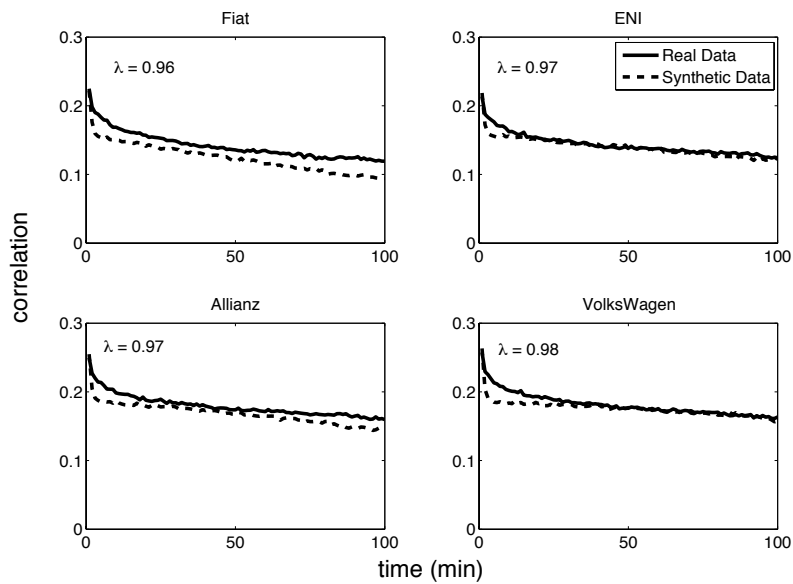


Figure 5. Autocorrelation functions of real data (solid line) and synthetic (dashed line) time series for the analyzed stocks.

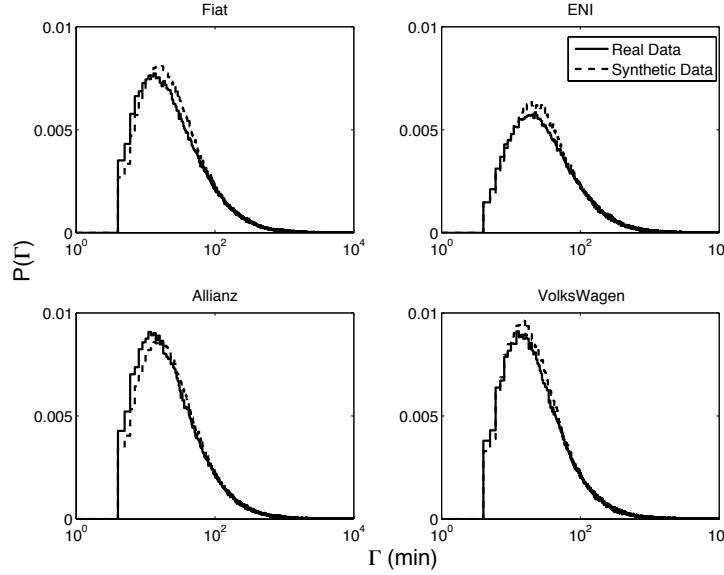


Figure 6. First passage time distribution of real data (solid line) and synthetic (dashed line) time series for the analyzed stocks.

the first time. We will denote the fpt as $\Gamma_\rho(t)$. Then

$$\Gamma_\rho(t) = \min\{\tau \geq 0; S(t + \tau)/S(t) \geq \rho\}.$$

In [8] we have shown how to calculate analytically such distribution for a semi-Markov process then we will not repeat that here. Using the best values for λ for each stock Fiat, Eni, Allianz and VolksWagen and choosing a value $\rho = 1.005$ for all of them we compare in Figure 6 results for the first passage time distribution for each stock. It can be noticed that they are almost identical improving the results obtained for a simple semi-Markov process presented in [8].

The results obtained here improve those obtained in our previous work [9, 8] even further showing that the semi-Markov approach is adequate to model high frequency financial time series.

4. Conclusions

We have modeled financial price changes through a semi-Markov model where we have added a weighted index. Our work is motivated by two main results: the existence in the market of periods of low and high volatility and our previous work [9], where we showed that an indexed semi-Markov model, is able to capture almost all the correlation in the square of returns present in real data. The results presented here show that the semi-Markov kernel is influenced by the past volatility and that its influence decreases exponentially with time. In fact, if the past volatility is used as an exponentially weighted index, the model is able to reproduce almost exactly the behavior of market

returns: the returns generated by the model are uncorrelated while the square of returns present a long range correlation very similar to that of real data.

We have also shown, by analyzing different stocks from different markets (Italian and German), that results do not depend on the particular stock chosen for the analysis even if the value of the weights may depends on the stock.

We stress that our model is very different from those of the ARCH/GARCH family. We do not model directly the volatility as a correlated process. We model returns and by considering the semi-Markov kernel conditioned by a weighted index the volatility correlation comes out freely.

References

- [1] F. Mainardi, M. Raberto, R. Gorenflo, E. Scalas. *Fractional calculus and continuous-time finance II: The waiting-time distribution*, Physica A, **287**, (2000), 468-481.
- [2] M. Raberto, E. Scalas, F. Mainardi. *Waiting-times and returns in high-frequency financial data: An empirical study*, Physica A **314**, (2002) 749-755.
- [3] G. D'Amico, J. Janssen, R. Manca. *Homogeneous semi-Markov reliability models for credit risk management*, Decisions in Economics and Finance, **28**, (2005), 79-93.
- [4] G. D'Amico, J. Janssen, R. Manca. *European and American options: The semi-Markov case*, Physica A, **388**, (2009), 3181-3194.
- [5] D.S. Silvestrov, F. Stenberg. *A pricing process with stochastic volatility controlled by a semi-Markov process*, Communications in Statistics: Theory and Methods, **33**, (2004), 591-608.
- [6] D.M. Guillaume, M.M. Dacorogna, R.R. DavVe, J.A. MWuller, R.B. Olsen, O.V. Pictet. *From the birds eye to the microscope: A survey of new stylized facts of the intra-daily foreign exchange markets*, Finance and Stochastics **1**, (1997), 95-129.
- [7] F. Petroni, M. Serva. *Spot foreign exchange market and time series*, The European Physical Journal B **34**, (2003), 495-500.
- [8] G. D'Amico, F. Petroni. *A semi-Markov model for price returns*, Physica A 2012, DOI: 10.1016/j.physa.2012.05.040.
- [9] G. D'Amico, F. Petroni. *A semi-Markov model with memory for price changes*, Journal of Statistical Mechanics: Theory and Experiment, P12009, 2011
- [10] N. Limnios, G. Oprisan. *An introduction to Semi-Markov Processes with Application to Reliability*, In D.N. Shanbhag and C.R. Rao, eds., Handbook of Statistics, **21**, (2003), 515-556.
- [11] G. D'Amico, F. Petroni, F. Prattico *First and second order semi-Markov chains for wind speed modeling*, Submitted, 2012.
- [12] G. D'Amico, *Age-usage semi-Markov models*, Applied Mathematical Modelling, **35**, (2011), 4354-4366.
- [13] J. Janssen and R. Manca, *Applied semi-Markov processes*, Springer, New York, 2006.
- [14] Barbu, V. and Limnios, N. (2008). *Semi-Markov Chains and Hidden Semi-Markov Models Toward Applications*. Springer-Verlag New York Inc.
- [15] I. Simonsen. M.H. Jensen, A. Johansen, *Optimal Investment Horizons*, Eur. Phys. J. **27** (2002) 583.
- [16] M. H. Jensen, A. Johansen, F. Petroni, I. Simonsen. *Inverse Statistics in the Foreign Exchange Market*. Physica A **340**, (2004) 678.